Computational Learning Theory	Fall Semester, 1999/00

Lecture 4: November 11

Lecturer: Yishay Mansour | *Scribe: Yael Bdolah, Dror Livnat*

4.1 Finite Horizon

In lecture number 3 we introduced the optimality equations:

$$U_t(h_t) = \max_{a \in A}\{r_t(s_t, a) + \sum_{j \in S} P_t(j|s_t, a)U_{t+1}(h_t, a, j)\},$$

Where $U_N(h_N) = r_N(s_N)$ for $h_N = (h_{N-1}, a_{N-1}, s_N)$.
We showed that there is an optimal policy which is a deterministic optimal policy. Next we show that there is an optimal policy that is both determinstic and Markovian.

4.1.1 Markovian Policy

Theorem 4.1 *Let U_t^* be a solution to the optimality equations then*

1. *For any t, $1 \le t \le N$, $U_t^*(h_t)$ depends on the history h_t only through the last state s_t.*
2. *There exist an optimal policy which is a Markovian deterministic policy.*

Proof: We will use a reversed induction to prove (1).
Induction Basis: $U_N^*(h_N) = r_N(s_N)$, therefore $U_N^*(h_N) = U_N^*(s_N)$
Induction Step: We assume the validity of the induction hypothesis for any n, $n \ge t+1$ and will prove the validity for $t = n$.

$$\begin{aligned} U_t^*(h_t) &= \max_{a \in A}\{r_t(s_t, a) + \sum_{j \in S} P_t(j|s_t, a)U_{t+1}^*(h_t, a, j)\} \\ &= \max_{a \in A} \underbrace{\{r_t(s_t, a) + \sum_{j \in S} P_t(j|s_t, a)U_{t+1}^*(j)\}}_{w(s,a)} \end{aligned}$$

Note that $w(s, a)$ depends merely on s and a. Therefore $U_t^*(h_t)$ depends solely on s_t. Thus,

$$U_t^*(h_t) = U_t^*(s_t)$$

To prove (2), let π be a Markovian deterministic policy that sutisfies:

$$\pi_t(s_t) = argmax_{a\in A}\{r_t(s_t,a) + \sum_{j\in S} P_t(j|s_t,a)U^*_{t+1}(j)\}$$

Since the policy's definition depends solely on s_t, namely the current state, π_t is a Markovian policy. □

Summary

Theorem 3.2 and theorem 4.1 lead to

$$V^*_N(s) = \max_{\pi\in\Pi^{HR}}\{V^\pi_N(s)\} = \max_{\pi\in\Pi^{MD}}\{V^\pi_N(s)\}$$

Namely, the optimal policy can always be chosen out of the group of Markovian deterministic policies.

4.1.2 An Algorithm for Constructing an Optimal Policy

In this section we develop an optimal Markovian deterministic policy. As shown in the previous section, by this we achieve a general optimal policy. The algorithm construction is done from $t = n$ back to $t = 1$.
The algorithm:

1. Let $t \leftarrow N$, $\forall s_N, U_N(s_N) = r_N(s_N)$
2. For $t = N - 1 downto 1 Do$

$$U_t(s_t) = \max_{a\in A}\{r_t(s_t,a) + \sum_{j\in S} P_t(j|s_t,a)U_{t+1}(j)\}$$

   and the optimal group of actions is:

$$A^*_{s_t}(t) = argmax_{a\in A}\{r_t(s_t,a) + \sum_{j\in S} P_t(j|s_t,a)U_{t+1}(j)\}$$

Any policy, π^*, satisfying $\forall t, \forall s_t, \pi^*_t(s_t) \in A^*_{s_t}$ is an optimal policy.

Computational Complexity

Let $K = |S|$ and $L = |A|$, then the time complexity of the described algorithm is $O(NLK^2)$. The latter is derived from repeating the iterative step for $t = 1, 2, ..., N$, calculating $U_t(s_t)$ for K states, computing the max for each of the L actions, and each computation may include all the K states.

4.1.3 Example: The Recruiting Problem

The Problem

A manager has to recruit a new employee and he can serially interview a finite group of candidates. There is a total order defined on the candidates' fitness to the opening, and there are no two employees with the same skill level. The manager is able to sort the candidates' fitness level after a short interview. After each interview the manager has two alternatives:

- recruit the currently interviewed candidate.
- continue to the next interview (and give up the chance of recruiting the previous candidate).

The Goal

To maximize the probability of recruiting the best candidate.

The Solution

We will first construct a corresponsing MDP for the problem, and then construct the optimal policy for it.
Let $A = \{Continue, QuitAndHire\}$ the possible actions, where $Continue$ stands for 'continue' and $QuitAndHire$ for 'quit and recruit'. Let $S = \{MaxSoFar, Other, 1, 0\}$ be the group of states of the MDP, standing for:

- MaxSoFar - the last interviewed candidate was the best so far.
- Other - the last interviewed candidate was NOT the best so far.
- 1 - the last interviewed candidate was THE best candidate in the entire group and was hired.
- 0 - the last interviewed candidate was NOT the best candidate in the entire group and was hired.

Figure 4.1 shows the resultant MDP, using the following transition probabilities:

- $q_t = Prob[\max\{x_1, ..., x_t\} = \max\{x_1, ..., x_N\}] = \frac{t}{N}$
- $r_t = Prob[x_{t+1} \geq \max\{x_1, ..., x_t\}] = \frac{1}{t+1}$

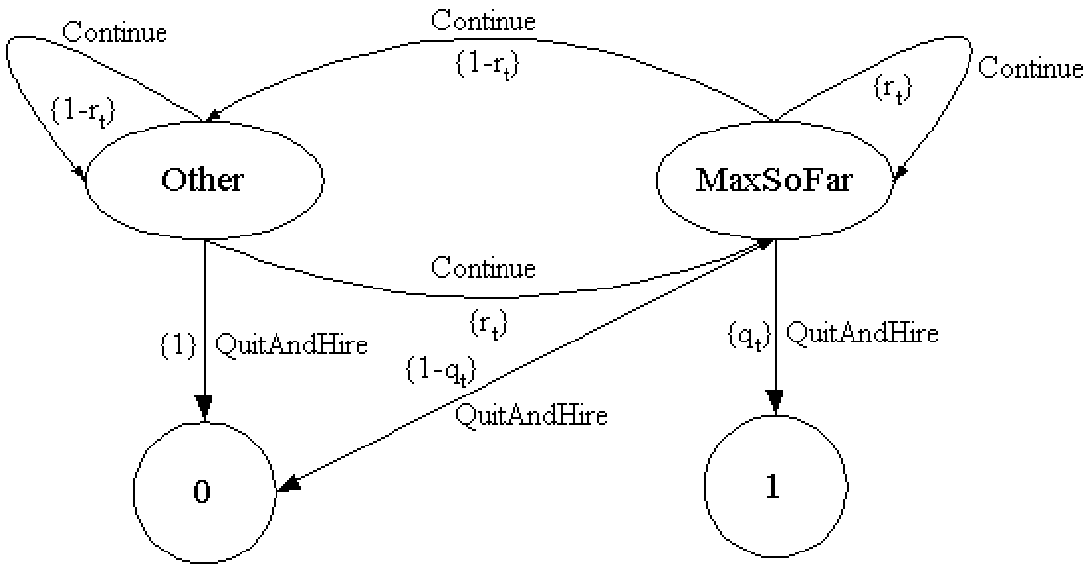

Figure 4.1: The recruiting problem MDP

where N is the finite time horizon (the size of the candidates group) and t is the current time (the number of candidates interviewed so far).
Writing the optimality equations for this problem we get:

$$U_t^*(1) = 1, \;\; U_t^*(0) = 0, \;\; U_{N+1}^*(MaxSoFar) = U_{N+1}^*(Other) = 0$$

$$\begin{aligned}
U_t^*(Other) &= \max\{0, r_t U_{t+1}^*(MaxSoFar) + (1-r_t)U_{t+1}^*(Other)\} \\
&= r_t U_{t+1}^*(MaxSoFar) + (1-r_t)U_{t+1}^*(Other) \\
U_t^*(MaxSoFar) &= \max\{q_t U_{t+1}^*(1), r_t U_{t+1}^*(MaxSoFar) + (1-r_t)U_{t+1}^*(Other)\} \\
&= \max\{q_t \cdot 1, r_t U_{t+1}^*(MaxSoFar) + (1-r_t)U_{t+1}^*(Other)\} \\
&= \max\{q_t, U_t^*(Other)\}
\end{aligned}$$

Assigning the values for q_t and r_t we get:

$$U_t^*(Other) = \frac{1}{t+1}U_{t+1}^*(MaxSoFar) + \frac{t}{t+1}U_{t+1}^*(Other) \tag{4.1}$$

$$U_t^*(MaxSoFar) = \max\{\frac{t}{N}, U_t^*(Other)\} \tag{4.2}$$

An optimal decision rule has the following properties:

- In state $Other$, always choose action $Continue$ (otherwise the return would to be zero).
- In state $MaxSoFar$, choose action $Continue$ if $\frac{t}{N} < U_t^*(Other)$ and action $QuitAndHire$ otherwise (this policy maximizes $U_t^*(MaxSoFar)$).

We now show that the optimal policy is of the form:

- Interview τ candidates, performing $a_t = Continue,\ t \leq \tau$.
- Quit and hire the first candidate after time τ, which is the best so far.

Optimality Proof

We show that the described policy agrees with the optimality equations.
We start by showing that if there is a time τ such that $\frac{\tau}{N} < U_\tau^*(MaxSoFar)$ then $\forall t, t < \tau$ we get $\frac{t}{N} < U_t^*(MaxSoFar)$. That is, if there exists a time $t = \tau$ in which it is preferred to *Continue*, then for any earlier time t it is also preferred to *Continue*. Later we show that such a time, τ, does exist.
Let $\frac{t}{N} < U_t^*(MaxSoFar)$, then according to equation 4.2 $U_\tau^*(MaxSoFar) = U_\tau^*(Other)$. Performing backword induction steps, and using equations 4.1 and 4.2 we show that for $t < \tau$ the inequality remains true:
$U_{\tau-1}^*(Other) = \frac{1}{\tau}U_\tau^*(MaxSoFar) + \frac{\tau-1}{\tau}U_\tau^*(Other) = U_\tau^*(Other) > \frac{\tau}{N}$
$U_{\tau-1}^*(MaxSoFar) = \max\{\frac{\tau-1}{N}, U_{\tau-1}^*(Other)\} > \frac{\tau}{N} > \frac{\tau-1}{N}$
We now show that for $t > \tau$ and $s = MaxSoFar$, it is preferred to *QuitAndHire*.
Let $t > \tau$, then according to the first half of the proof,
$U_t^*(MaxSoFar) = \frac{t}{N}$
and,

$$
\begin{aligned}
U_t^*(Other) &= \frac{1}{t+1}U_{t+1}^*(MaxSoFar) + \frac{t}{t+1}U_{t+1}^*(Other) \\
&= \frac{1}{N} + \frac{t}{t+1}U_{t+1}^*(Other) \\
&= \frac{1}{N} + \frac{t}{t+1}\frac{1}{N} + \frac{t}{t+2}\frac{1}{N} + \ldots \\
&= \frac{1}{N}\sum_{i=0}^{N-t}\frac{t}{t+i} \sim \ln(\frac{N}{t})
\end{aligned}
$$

We conclude this proof by showing that such a τ exists, that is, for $N \geq 2$ there exists $\tau \geq 1$ that meets the above requiremints. Let us assume $\tau = 0$ we get:

$U_1^*(Other) = \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{N} > \frac{1}{2} \geq \frac{1}{N} = U_1^*(MaxSoFar) \geq U_1^*(Other)$
which is a circular inequality, and thus leads to cnotradiction.
Note that for $N \geq 2$ we always get $\tau \geq 1$:
$U_1^*(Other) = U_1^*(MaxSoFar) = ... = U_\tau^*(Other) = U_\tau^*(MaxSoFar)$
and for $t > \tau$ we have,
$U_t^*(MaxSoFar) = \frac{t}{N}$
$U_t^*(Other) = \frac{t}{N}(\frac{1}{t} + \frac{1}{t+1} + ... + \frac{1}{N-1})$.
We therefore choose $Continue$ if $\frac{1}{t} + ... + \frac{1}{N-1} > 1$ and $QuitAndHire$ otherwise.

A Numeric Example

Let $N = 5$, we get $\frac{1}{3} + \frac{1}{4} < 1$ and $\frac{1}{2} + \frac{1}{3} + \frac{1}{4} > 1$ and therefore we choose $\tau = 2$.

An Asymptotic Example

Let $N \to \infty$, we get

$$\sum_{j=\tau}^{N-1} \frac{1}{j} \sim \ln \frac{N}{\tau}.$$

We therefor shearch for τ such that
$\ln \frac{N}{\tau+1} < 1$ and $\ln \frac{N}{\tau} > 1$,
which leads to approximately

$$\tau \sim \frac{N}{e}$$

Conclusion

The best policy is to perform $\frac{N}{e}$ interviews, keeping in mind only the level of the best candidate so far. Then, starting at candidate number $\frac{N}{e} + 1$, $QuitAndHire$ on the first candidate which satisfies the requirement of $BestSoFar$. This way the asysmptotic success probability is $\frac{\tau}{N} = \frac{1}{e}$.

4.2 Infinite Horizon Problems

4.2.1 The Return Function

We introduce three popular return functions for the infinite horizon problem:

1. The **expected sum of the immediate rewards**, i.e.

$$\begin{aligned} V^\pi(s) &= \lim_{N\to\infty} E_s^\pi[\sum_{t=1}^{N} r_t(X_t, Y_t)] \\ &= \lim_{N\to\infty} V_N^\pi(s) \end{aligned}$$

   Note that this return function may diverge (for examle, if the immediate expected reward is non-zero).

2. The **expected discounted sum of the immediate rewards**, i.e.

$$V_\lambda^\pi(s) = \lim_{N\to\infty} E_s^\pi[\sum_{t=1}^{N} \lambda^{t-1} r_t(X_t, Y_t)], \quad 0 < \lambda < 1$$

   In this case, a sufficient condition for convergence is: $|r(\cdot,\cdot)| \le M$
   Under this condition we can find an upper bound to the return function:

$$V_\lambda^\pi(s) \le \sum_{t=1}^{N} \lambda^{t-1} M = \frac{M}{1-\lambda}$$

   Note that this bound is very sensative to the value of the paramter λ. Also, we can get the previous return function as follows:

$$\lim_{\lambda\to 1} V_\lambda^\pi(s) = V^\pi(s)$$

3. The **expected average reward**

$$\begin{aligned} g^\pi(s) &= \lim_{N\to\infty} \frac{1}{N} E_s^\pi[\sum_{t=1}^{N} r(X_t, Y_t)] \\ &= \lim_{N\to\infty} \frac{1}{N} V_N^\pi(s) \end{aligned}$$

   This limit does not always exist. A sufficient condition for the limit to exist is:

   (a) S is finite.
   (b) π is Markovian and stationary.
   (c) the MDP with π is non-periodic.

   These conditions will be discussed further in a later lecture.

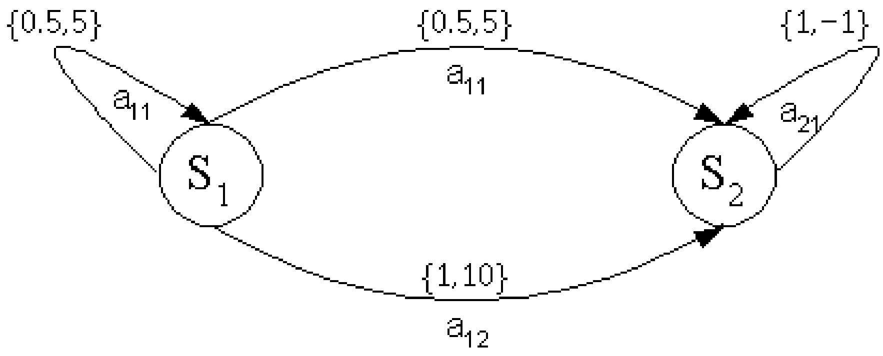

Figure 4.2: Infinite horizon example

4.2.2 Example 1

This example is an expantion of exmaple 2 using two states, given in lecture 3.
We first examine the value gathered from different return functions using two specific policies:

1. π_1 - always chooses a_{11} when in state s_1

2. π_2 - always chooses a_{12} when in state s_1

Let us start by calculating V_N^{π}:

$$\begin{aligned}
V_N^{\pi_2} &= 10-(N-2)=12-N \\
V_N^{\pi_1} &= 5+[\frac{1}{2}\cdot 5+\frac{1}{2}\cdot(-1)]+[\frac{1}{4}\cdot 5+\frac{3}{4}\cdot(-1)]+... \\
&= [10-\frac{1}{2^{N-2}}\cdot 5]-(N-2)+(1-\frac{1}{2^{N-2}}) \\
&= 13-N-\frac{6}{2^{N-2}}
\end{aligned}$$

For $N\to\infty$ the gap between the two policies goes to 1 in favor of π_1.
The three suggested return functions evaluate to:

1. The expected sum of the immediate rewards:

$$V^{\pi_1}(s_1)=V^{\pi_2}(s_1)=-\infty$$

2. Expected average reward:

$$\begin{aligned} g^{\pi_2}(s_1) &= \lim_{N\to\infty} \frac{12-N}{N} = -1 \\ g^{\pi_1}(s_1) &= -1 \end{aligned}$$

3. Expected discounted sum:

$$\begin{aligned} V_\lambda^{\pi_1}(s_1) &= 5 + \lambda[\frac{1}{2}\cdot 5 + \frac{1}{2}\cdot(-1)] + \lambda^2[\frac{1}{4}\cdot 5 + \frac{3}{4}\cdot(-1)] + ... \\ &= \frac{5}{1-\frac{\lambda}{2}} - \sum_{i=1}^{\infty}(1-\frac{1}{2^i})\lambda^i \\ &= \frac{5}{1-\frac{\lambda}{2}} - \frac{\lambda}{1-\lambda} + \frac{\frac{\lambda}{2}}{1-\frac{\lambda}{2}} \\ &= \frac{10+\lambda}{2-\lambda} - \frac{\lambda}{1-\lambda} = \frac{10-11\cdot\lambda}{(2-\lambda)\cdot(1-\lambda)} \\ V_\lambda^{\pi_2}(s_2) &= 10 + \sum_{i=1}^{\infty}\lambda^i(-1) = 10 - \frac{\lambda}{1-\lambda} \end{aligned}$$

4.2.3 The Expected Discounted Sum Return Function

Here are some possible explanations for the λ parameter.

1. In economical problems the λ parameter may be interpreted as the interest rate.

2. Consider a finite horizon problem where the horizon is random, i.e.

$$V_N^\pi(s) = E_s^\pi E_N[\sum_{i=1}^{N} r(X_t, Y_t)]$$

   assuming that the final value of all the states is equal to 0.

   Let N be distributed geometricly with parameter λ. The probability of generating a horizon of length n is:
   $Prob[N = n] = (1-\lambda)\lambda^{n-1}$

Lemma 4.2 *$V_N^\pi(s) = V_\lambda^\pi(s)$, assuming that $|r(\cdot,\cdot)| < M$*

Proof:

$$\begin{aligned}
V_N^\pi(s) &= E_s^\pi\{\sum_{n=1}^{\infty}[\sum_{t=1}^{n} r(X_t,Y_t)](1-\lambda)\lambda^{n-1}\} \\
&= E_s^\pi[\sum_{t=1}^{\infty} r(X_t,Y_t)(1-\lambda)\sum_{n=t}^{\infty}\lambda^{n-1}] \\
&= E_s^\pi[\sum_{t=1}^{\infty} r(X_t,Y_t)](1-\lambda)\frac{\lambda^{t-1}}{1-\lambda}] \\
&= E_s^\pi[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t,Y_t)] \\
&= V_\lambda^\pi(s)
\end{aligned}$$

□

If we look back at example 1 we could add to it an additional state, Λ, that behaves as a 'black hole', see figure 4.3. Once the system reaches this state it stays there forever, getting an immediate reward of value 0.
The probablity to move into state Λ is $(1-\lambda)$ from any state. All other probabilities given in the original example are multiplied by λ.
The sum of the immediate rewards from the new model is equal to the discounted sum of the immediate rewards from the original model.

4.2.4 Markovian Policy

We show that for every initial state, s, and a history dependent policy, π, there exists a Markovian policy π' such that the distribution on (X_t, Y_t) is equal for π and π'. We will later derive the return functions of π and π' and show their equality.

Theorem 4.3 *Let $\pi = (d_1, d_2, ...) \in \Pi^{HR}$.*
Then $\forall s \in S$ there exists a Markovian stochastic policy $\pi' = (d'_1, d'_2, ...) \in \Pi^{MR}$,
that sutisfies $Prob_{\pi'}[X_t = j, Y_t = a | X_1 = s] = Prob_\pi[X_t = j, Y_t = a | X_1 = s]$

Proof: For every $j \in S$ and $a \in A_j$ we define π' as follows:
$q_{d'_t(j)}(a) = Prob_\pi[Y_t = a | X_t = j, X_1 = s]$
We first show that this definition results in the same distibution over **actions**

$$\begin{aligned}
Prob_{\pi'}[Y_t = a | X_t = j] &= Prob_{\pi'}[Y_t = a | X_t = j, X_1 = s] \\
&= Prob_\pi[Y_t = a | X_t = j, X_1 = s]
\end{aligned}$$

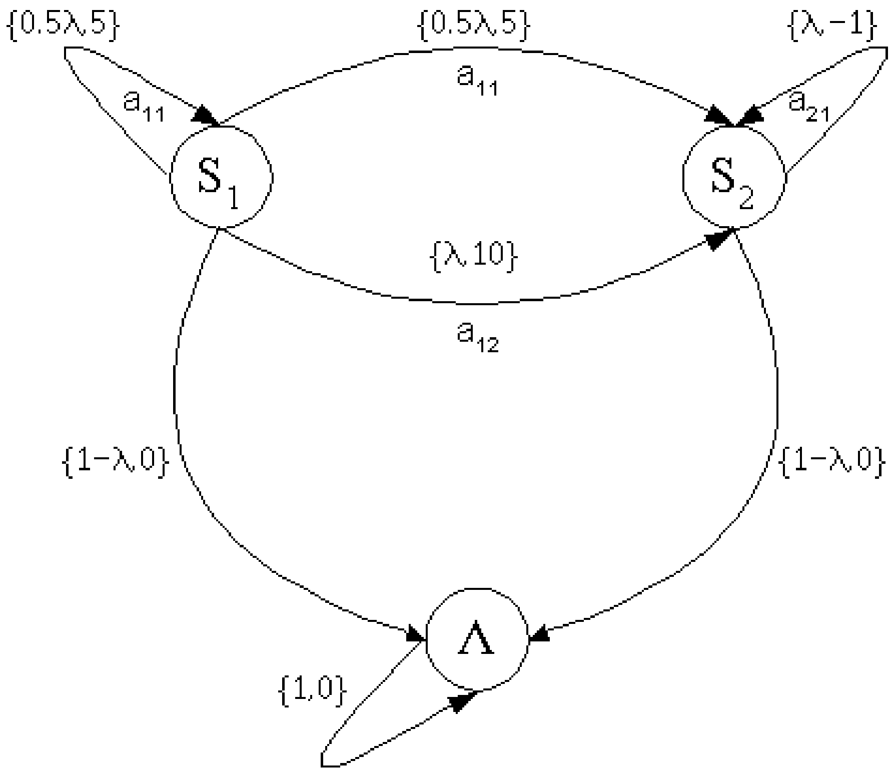

Figure 4.3: Implementation of a discounted return function using a non discounted return function

The first equality is derived from the fact that π' is Markovian. The second equality is by definition.

Next we show that the distribution of **states** is equal under π and π', i.e.,
$Prob_{\pi'}[X_t = j|X_1 = s] = Prob_{\pi}[X_t = j|X_1 = s]$.
We prove this part by an induction on t. The idea behind the proof of this part is that if at a certain step we have an equal distribution over the group of states, and we are taking the same stochastic action, we will endup with same distribution over the group of states.
Induction Basis: for $t = 1 \quad X_1 = s$ in π and in π'
Induction Step: We assume that the distribution over states and actions in π and in π' is

identical until the time $t-1$

$$\begin{aligned}
Prob_{\pi}[X_t = j|X_1 = s] &= \sum_{k\in S}\sum_{a\in A_k} Prob_{\pi}[X_{t-1} = k, Y_{t-1} = a|X_1 = s]p(j|k,a) \\
&= \sum_{k\in S}\sum_{a\in A_k} Prob_{\pi'}[X_{t-1} = k, Y_{t-1} = a|X_1 = s]p(j|k,a) \\
&= Prob_{\pi'}[X_t = j|X_1 = s]
\end{aligned}$$

Since $Prob_{\pi'}[X_t = j, Y_t = a|X_1 = s] = Prob_{\pi'}[X_t = j|X_1 = s] \cdot Prob_{\pi'}[Y_t = a|X_t = j, X_1 = s]$ this concludes this proof. $\square$

The Return Function

$$\begin{aligned}
V_N^{\pi}(s) &= \sum_{t=1}^{N-1}\sum_{j\in S}\sum_{a\in A_j} r(j,a)\cdot Prob[X_t = j, Y_t = a|X_1 = s] \\
&\quad + \sum_{j\in S}\sum_{a\in A_j} [r_N(j)]\cdot Prob[X_N = j, Y_N = a|X_1 = s]
\end{aligned}$$

Therefore:

1. $\forall N,\ V_N^{\pi}(s) = V_N^{\pi'}(s)$
   Since we proved in the last theorem that the dsitribution function is equal for π and π'.

2. $g^{\pi}(s) = g^{\pi'}(s)$
   Since (1) is true for all N.

3. $V_{\lambda}^{\pi}(s) = V_{\lambda}^{\pi'}(s)$

One should note that theorem 4.3 does not hold for $\pi \in \Pi^{HD}$ and $\pi' \in \Pi^{MD}$, since the random property of π' allows the modeling of all histories under one state.